Journal of Economics
and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq

مقارنة بين طريقتي الامكان الشبه الجزائيه و الامكان الشبه الحدي في تقدير معلمات الانموذج الثنائي المتعدد المستويات

الباحث/ زينب نهاد محمد الراوي
جامعة بغداد/ مركز الحاسبة الالكترونية
zainab.alrawi@yahoo.com

أ.م. ايمان حسن العاني
جامعة بغداد/ كلية الادارة والاقتصاد/ قسم الاحصاء
eman.alani1962@gmail.com

Received:5/5/2020 Accepted :21/6/2020 Published :October / 2020

هذا العمل مرخص تحت اتفاقية المشاع الابداعي نسب المُصنَّف - غير تجاري - الترخيص العمومي الدولي 4.0
Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

مستخلص البحث:

تعد النماذج متعددة المستويات من اهم النماذج الواسعة الاستعمال في تطبيق وتحليل البيانات التي تتصف بكون المشاهدات تتخذ الشكل الهرمي، وفي بحثنا هذا تم تناول انموذج الانحدار اللوجستي الثنائي متعدد المستويات (انموذج حد ثابت عشوائي وميل عشوائي)، هنا تبرز اهمية البحث ان نماذج الانحدار الاعتيادية تقوم بحساب التباين العام للأنموذج وعدم امكانيتها قراءة التباينات والتغيرات بين المستويات لكن في حالة نماذج الانحدار المتعددة المستويات فان حساب التباين الكلي لها يكون غير دقيق وعليه فأن هذه النماذج تقوم بحساب التباينات والتغيرات لكل مستوى من مستويات الانموذج حيث يهدف البحث الى تقدير معلمات هذا الانموذج باستعمال طرائق تقريبية طريقة الامكان الشبة الجزائية (Penalized quasi-likelihood)، طريقة الامكان الشبه الحدي (Marginal quasi-likelihood))، اذ تم استخدام اسلوب المحاكاة للمقارنة بين طرائق التقدير باختلاف حجوم العينات، ومن خلال معيار المقارنة متوسط مربعات الخطأ (MSE) للحصول على الطريقة الافضل في تقدير المعلمات، اظهرت النتائج التي تم الحصول عليها باستعمال اسلوب المحاكاة بأن طرائق التقدير اعطت نتائج متقاربة جدا الا ان طريقة الامكان الشبه الحدي (MQL) هي الافضل في جميع حجوم العينات.

المصطلحات الرئيسة للبحث: الانحدار اللوجستي، طريقة لابلاس، دالة الامكان شبه المتكاملة، خوارزمية فشر الرقمية، المربعات الصغرى الموزونة التكرارية.

البحث مستل من رسالة ماجستير.

1- المقدمة (Introduction)

انموذج الانحدار اللوجستي يعتبر من نماذج الانحدار اللاخطية ويوضح العلاقة بين المتغير التابع ثنائي الاستجابة والمتغيرات التوضيحية، حيث قام العديد من الباحثين باستعمال انموذج الانحدار اللوجستي وكان الباحث (verhulst) اول من استخدم دالة اللوجستك لوصف نمو المجتمع وتسمى بدالة النمو (growth function)، وفي عام 1920استخدم الباحثان (pearl&reed) الدالة لحساب نمو السكان واطلق عليها بدالة اللوجستك فيما بعد، ولكونه من النماذج الملائمة للبيانات الثنائية يتم استخدامه في كثير من المجالات المتعلقة بالعلوم الحياتية والزراعية والطبية وقد تناولنا في هذا البحث انموذج الانحدار اللوجستي متعدد المستويات [1][3]، ويعرف المستوى هو مجموعه من المفردات او الوحدات والتي تشترك في مجموعة من الصفات المتشابهة والتي تميزها عن مجموعة اخرى ويمكن التعبير عنه بانه مجتمع احصائي فريد يمكن تحديد ملامحه والوصول اليه بذاته وفي حالة وجود مستوى اعلى (ثاني) لهذا المستوى فلابد ان يكون هذا المستوى عبارة عن وحدات اكبر وتشمل المفردات او الوحدات في المستوى السابق، مثلا المدارس يتم دمج الطلاب (المستوى الاول) في الفصول (المستوى الثاني) والفصول الدراسية متداخلة في المدارس (المستوى الثالث). ان اسلوب تحليل متعدد المستويات يقوم بتقسيم مجتمع الدراسة الى عدد من المستويات حيث يمثل مستوى الاول غالبا الافراد المستهدفين اما بقية المستويات تعتمد على موضوع الدراسة فمثلا اذا كان موضوع الدراسة خاص بالجانب الصحي فسيمثل المستوى الثاني المستشفيات والمستوى الثالث يمثل المحافظة ومن هنا تبرز مشكلة البحث اذ ان نماذج الانحدار الاعتيادية تقوم بحساب التباين العام للأنموذج وعدم امكانيتها قراءة التباينات والتغيرات بين المستويات لكن في حالة نماذج الانحدار المتعدد المستويات فان حساب التباين الاجمالي لها يكون غير دقيق وعليه فأن هذه النماذج تقوم بحساب التباينات والتغيرات لكل مستوى من مستويات الانموذج، وان الهدف الاساسي للبحث هو تقدير معلمات انموذج الانحدار اللوجستي متعدد المستويات باستعمال طريقة الامكان الشبه الجزائية (Penalized quasi-likelihood) وطريقة الامكان الشبه الحدي(Marginal quasi-likelihood)، ومن ثم المقارنة بين طريقتي التقدير من خلال معيار المقارنة متوسط مربعات الخطأ (MSE) للوصول الى الطريقة الافضل في التقدير[6][7].

يتضمن البحث توضيح حول انموذج الانحدار اللوجستي الثنائي متعدد المستويات وكذلك طرائق التقدير ومن ثم دراسة المحاكاة التي تتضمن خطوات دراسة المحاكاة ونتائج دراسة المحاكاة وفي نهاية البحث يتضمن الاستنتاجات.

2- انموذج الانحدار اللوجستي (logistic Regression Model)

يعرف انموذج الانحدار اللوجستي على انه احد نماذج الانحدار اللاخطية الذي تكون فيه العلاقة بين المتغير التابع (y) (متغير الاستجابة) والمتغيرات التوضيحية ($x_1, x_2, x_3, \ldots x_n$) غير خطية، وان الانحدار اللوجستي يبنى على فرض اساسي وهو ان المتغير التابع (y) ثنائي الاستجابة يأخذ أحدى القيمتين (1,0) اما نجاح اي حدوث الاستجابة (Success) بأحتمال μ_i او فشل عدم حدوث الاستجابة (Failure) بأحتمال ($1-\mu_i$) لذلك فأن المتغير التابع y_i يتوزع توزيع برنولي[4].

اي ان :-

$$Y_i \sim Ber(\mu_i)$$
$$i=1,2,\ldots..n$$

وان دالة الكثافة الاحتمالية تكون بالصيغة الاتية :-

$$P_r(Y=y_i) = \mu_i^{y_i} (1-\mu_i)^{1-y_i} \qquad (1)$$
$$y_i = 0,1$$

اذ ان :-

y_i : متغير تابع ثنائي الاستجابة.

μ_i احتمال حدوث الاستجابة عندما $yi = 1$.

$1-\mu_i$:احتمال عدم حدوث الاستجابة عندما $yi = 0$.

لذلك فأن توقع المتغير yi يمثل احتمال حدوث الاستجابة (μ_i)

$$E(y_i) = P_r(Y=1) = \mu_i \qquad (2)$$

اما بالنسبة لتباين المتغير y_i حسب توزيع برنولي:

$$V(y_i) = \mu_i\,(1-\mu_i) \qquad (3)$$

ليكن $(x_1, x_2, x_3, \dots \dots x_k)$ عبارة عن مجموعة من المتغيرات التوضيحية و n تمثل عدد المشاهدات لهذه المتغيرات التي تكون المصفوفة التالية:-

$$X = \left(x_{iq}\right)_{n*k} \qquad (4)$$

اذ ان :-

X :تمثل مصفوفة المتغيرات التوضيحية.

$i = 1, 2, 3, \dots\dots n$; n: تمثل عدد المشاهدات (حجم العينة).

$q = 1, 2, 3, \dots\dots k$; k: تمثل عدد المتغيرات التوضيحية.

حيث ان: $y_i = [y1, y2, y3 \dots\dots, yn]$ تمثل عينة عشوائية من المتغير ثنائي الاستجابة وان $y_i \in \{0, 1\}$.

وبالتالي يمكن التعبير عن انموذج الانحدار اللوجستي بالصيغة التالية:-

$$y_i = \mu_i + \varepsilon i \qquad (5)$$

حيث ان :-

μ_i :تمثل دالة الانحدار اللوجستي (دالة الاستجابة اللوجستية).

$$\mu_i = p(y=1) = \frac{e^{xi'\beta}}{1+e^{xi'\beta}} \qquad (6)$$

β :متجه المعلمات ابعاده $(P\times 1)$.

$x_i = \{x_{i1}, x_{i2} \dots x_{ip}\}$: متجه صفي من المتغيرات التوضيحية ابعاده $(1\times p)$.

εi :تمثل الخطأ العشوائي.

$$\varepsilon_i = y_i - \mu_i$$

حيث ان حد الخطأ له متوسط يساوي صفر وتباين مساوياً الى تباين المتغير التابع وكالاتي:-

$$E(\varepsilon_i) = E(y_i) - E(\mu_i) = \mu_i - \mu_i = 0 \qquad (7)$$

$$V(\varepsilon_i) = V(y_i) = \mu_i(1-\mu_i) \qquad (8)$$

اي ان حد الخطأ يكون له متوسط يساوي صفر وتباين $\mu_i(1-\mu_i)$ ونلاحظ ان تباين حد الخطأ يعتمد على قيم احتمال الاستجابة μ_i اي على قيم المتجه x_i وبالتالي سوف يكون تباين حد الخطأ غير متجانس[11] .

3- انموذج الانحدار اللوجستي الثنائي متعدد المستويات(انموذج القاطع والميل العشوائيين) (Multilevel Binary Logistic Regression)

اسلوب تحليل متعدد المستويات يقوم بتقسيم مجتمع الدراسة الى عدة مستويات، حيث ان المستوى الاول غالبا ما يمثل الافراد المستهدفين في الدراسة اما بقية المستويات فتعتمد على نوع الدراسة، وعندما يكون هيكل وطبيعة البيانات المستحصل عليها من مجتمع معين في شكل هرمي (Hierarchical) فأن العينة المسحوبة من مثل هذه المجتمعات يمكن اعتبارها كعينة متعددة المراحل حيث ان عامل الوقت والكلفة يمكن ان تأخذ بنظر الاعتبار في مثل هذه العينات، تعطي هذه العينات وصف كمي لمتغيرات المجتمع لكن هناك الكثير من الصعوبات للوصول الى نتائج عن تلك المتغيرات، اذا ان البيانات تعتمد على بعضها البعض ويكون هناك حالة من الارتباط بين المشاهدات بسبب تعدد مستوياتها ضمن البيانات الهرمية ففي هذه الحالة فأن استخدام النماذج الاحصائية وحيدة المستوى غير فعال ولذلك اننا بحاجة الى انموذج احصائي معقد يسع ذالك الشكل الهرمي للبيانات للوصول الى استدلالات جيدة عن المجتمع[9].

ويطلق عليه ايضا انموذج الانحدار اللوجستي الهرمي (Hierarchical Logistic Regression) او انموذج الانحدار اللوجستي ذو التأثيرات العشوائية وهو نموذج يستخدم في حالة البيانات الهرمية التي تتكون من مستويات متداخلة بحيث تتأثر المتغيرات في كل مستوى بالمستويات الاخرى ،وعادة يكون الهدف من استخدام نماذج الانحدار اللوجستية متعددة المستويات هو فحص البيانات الهرمية وايجاد العلاقة بين متغيرين او اكثر احدهما معتمد ينتمي الى فئة المتغيرات الوصفية الثنائية والتي لها توزيع برنولي في حين تتنوع طبيعة المتغيرات التوضيحية المتوقع ان تكون لها علاقة بذلك المتغير المعتمد ومن خلال اجراءات تقدير معلمات الانموذج يمكن تحديد طبيعة تلك العلاقة بين المتغير المعتمد والمتغيرات التوضيحية في المستويات المختلفة للبيانات، ويعرف الانموذج بالصيغة التالية[12][3]:-

$$y_{ij} = m_{ij} + \varepsilon_{ij}$$
$$logit(m_{ij}) = \beta_{oj} + \beta_{1j}x_{ij} + \beta_{2j}x_{ij} \qquad (9)$$

حيث ان :-

y_{ij} (متغير الاستجابة الثنائي للوحدة i ضمن المجموع (العنقود) j).

المعادلة (11) تمثل المستوى الاول.

$(B_{oj}, B_{1j}, \beta_{2j})$:تمثل معاملات المستوى الاول.

ويتم صياغه انموذج فرعي لكل معامل في المستوى الثاني كالاتي:-

$$B_{0j} = \alpha_{oo} + \alpha_{o1}z_j + b_{oj}$$
$$B_{1j} = \alpha_{10} + \alpha_{11}z_j + b_{1j} \qquad (10)$$
$$\beta_{2j} = \alpha_{20} + \alpha_{21}z_j + b_{2j}$$

ومن خلال تعويض معادلة المستوى الثاني (12) بمعادلة المستوى الاول (11) نحصل على :-

$$logit\,(\mu_{ij}) = (\alpha_{oo} + \alpha_{o1}z_j + b_{oj}) + (\alpha_{1o} + \alpha_{11}z_j + b_{1j})x_{ij} + (\alpha_{20} + \alpha_{21}z_j + b_{2j})x_{ij}$$

وبإعادة ترتيب المعادلة:-

$$= (\alpha_{oo} + \alpha_{01}z_j + \alpha_{10}x_{ij} + \alpha_{11}z_jx_{ij} + \alpha_{20}x_{ij} + \alpha_{21}z_jx_{ij} + b_{oj} + b_{1j}x_{ij} + b_{2j}x_{ij}) \qquad (11)$$

وان :-

$$i = 1, 2, 3 \ldots n_j$$
$$j = 1, 2, 3 \ldots M$$

N : تشير الى عدد المجاميع.

n_j: تشير الى عدد المشاهدات داخل كل مجموعة.

وفي هذا النموذج يتم افتراض ان عدد المفردات في المستوى الاول مساوي لكل المجموعات

x_{ij} :قيمة المتغير التوضيحي للمفردة i في المجموعة j (متغير التوضيحي في المستوى الاول).

z_j :قيمة المتغير التوضيحي على مستوى المجموعة j(متغير التوضيحي للمستوى الثاني).

$\alpha_{oo}, \alpha_{01}, \alpha_{1o}, \alpha_{20}, \alpha_{21}$: معاملات المستوى الثاني وتسمى ايضا التاثيرات الثابتة.

b_{oj}, b_{1j}, b_{2j} :تأثيرات العشوائية للمستوى الثاني.

وان :-

$$\begin{bmatrix} b_{0j} \\ b_{1j} \\ b_{2j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_{11} & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_{22} \end{bmatrix}\right)$$

وان [8pp3] :-

$[\alpha_{oo} + \alpha_{01}z_j + \alpha_{10}x_{ij} + \alpha_{11}z_jx_{ij} + \alpha_{20}x_{ij} + \alpha_{21}z_jx_{ij}]$:يمثل جزء الثابت للأنموذج.

$[b_{oj} + b_{1j}x_{ij} + b_{2j}x_{ij}]$:يمثل الجزء العشوائي للأنموذج.

4- طرائق التقدير (Estimation methods)

ينتمي النموذج (11) الى نماذج الخطية الهرمية المعممة ولكتابة الارجحية لنموذج الميل العشوائي، ليكن المتجه y يتكون من العناصر y_{ij} هذه العناصر مشروطة بالتاثيرات العشوائية والتي من المفترض ان تكون مستقلة عن بعضها البعض ولكل عنصر كثافة مشروطة:-

$$f(y_{ij}|b_j(\mathrm{y_{ij}}|\mathrm{b_j}) \sim Bernoulli\ (\mu_{ij}) \qquad (12)$$

وان القيمة المتوقعة لتوزيع برنولي والتي تساوي (μ_{ij}) فأنها بعد تطبيق دالة الربط المحددة تتنمذج كدالة خطية للمتغيرات التوضيحية، بعد تحديد توزيع التأثيرات العشوائية حيث ان الافتراض النموذجي بانه:

$$b_1, \ldots b_M \sim N(0, \Sigma) \qquad (13)$$

ومن اجل تقدير معلمات مثل هذه النماذج $(HGLM)$ عادة ما نستخدم تقدير الامكان الاعظم الحدي وفي هذه طريقة نحصل على الامكان الحدي للبيانات المشاهدة والتي يتم الحصول عليها من خلال تكامل توزيع التأثيرات العشوائية، ومن خلال معادلة (12) و (13)يمكن كتابة الارجحية الحدية (L) المشروطة بالمتغيرات التوضيحية :-

$$L(y) = \int \prod_{j=1}^{M} \prod_{i=1}^{n_j} f y_{ij} | b_j (y_{ij} | b_j) f_{bj}(b_j)\, db_j$$
$$= \prod_{j=1}^{M} \int \prod_{i=1}^{n_j} f y_{ij} | b (y_{ij} / b_j) f_b(b)\, db \qquad (14)$$

حيث ان L(y) تعتمد على المعلمات $\alpha_{00}, \alpha_{10}, \alpha_{01}, \alpha_{11}, \alpha_{20}, \alpha_{21}, \sigma_{20}, \sigma_{21}, \sigma_{00}, \sigma_{10}, \sigma_{11}$
وان احتساب المعادلة (14) يتطلب حساب N من التكاملات اضافة الى تعظيمها نسبة الى المعلمات المجهولة في الإنموذج , عموما المعادلة (14) ليس لها حل مغلق ولايمكن حلها بالطرائق التحليليه الاعتيادية لذلك يجب حلها بالطرائق العددية التكرارية باستخدام بعض الخوارزميات مثل خوارزمية (EM) وخوارزمية نيوتن رافسون (NR) وتم تطوير عدة طرق فعاله لحساب وتعظيم الارجحية طريقة شبه الارجحية وتكامل مونتي كارلو و التربيع العددي وفي هذا البحث سنعرض بعض طرائق تقدير معلمات الانحدار اللوجستي متعدد المستويات بهدف ايجاد مقدرات لمعلمات الانموذج والتي تتصف بصفــــات جيدة تؤهلنــــا لتكوين انموذج تقديــر يتم الاعتماد عليه في اغراض مختلفة والوصول الى نتــائج اكثر دقــة[6][3]:

- طريقة الامكان الشبة الجزائية Method Penalized quasi-likelihood.
- طريقة الامكان الشبه الحدي Method Marginal quasi- likelihood.

1-4: طريقة الامكان الشبه الجزائية Penalized quasi-likelihood (pql)

اقترحت طريقة الامكان الشبه الجزائية من قبل Green عام 1987 كطريقة للاستدلال التقريبي للنماذج متعددة المستويات والتي تستخدم خوارزميات المربعات الصغرى الموزونة لتقدير المعلمات ، اذ تميل طريقة (pql) الى تقليل الى حد ما من مكونات التباين والتأثيرات الثابتة بالقيمة المطلقة عندما يتم تطبيقها على البيانات الثنائية العنقودية. وكما ذكرنا سابقا ان المعادلة (14) ليس لها حل مغلق ولذلك يتم تقدير المعلمات بالاعتماد على معادلة شبه الاحتمال متكاملة ومن ثم تقريب لابلاس وكمــــــا يلي[8][5][2]:-
ولنفرض لدينا متجة التأثيرات العشوائية b لـ q من الابعاد، y_{ij}مستقلة بشكل مشروط مع الوسط والتباين

$$E(y_{ij}/b) = \mu_{ij}^{b}$$
$$Var(y_{ij}/b) = \varnothing a_{ij} v(\mu_{ij}^{b})$$

وان :-
$v(.)$:تباين الدالة المحددة.
a_{ij} :ثابت معلوم.
ø :معلمة التشتت والتي قد تكون معلومة او مجهولة.
يرتبط المتوسط الشرطي μ_{ij}^{b} بالتنبوء الخطي $n_i^b = X_{ij}^t\alpha + Z_{ij}^t b$ بواسطة دالة الربط $g(\mu_{ij}^{b}) = n_{ij}^{b}$ بمعكوس $h=g^{-1}$، حيث ان α متجه للتأثيرات الثابتة من درجة $p \times 1$ وان المتوسط الشرطي يكون كما يلي :-

$$E(y/b) = h(X\alpha + Zb) \qquad (15)$$

حيث ان:-
y :يرمز لمتجة المشاهدات $y=(y_1, y_2, .. y_n)^t$.
α :متجه التاثيرات الثابتة $p\times 1$.
b :متجه التاثيرات العشوائية $q \times 1$.
وبافتراض ان b له توزيع طبيعي متعدد المتغيرات بمتوسط (0) ومصفوفة تباين مشترك$D(\theta)$.
حيث ان :-
θ :متجه غير معلوم من مكونات التباين.

ولتقدير قيمة α و θ سنستخدم تكامل دالة شبه الامكان (دالة الامكان شبة المتكاملة quasi-likelihood function والتي تعرف بالصيغة التالية :-

$$e^{ql(\alpha,\theta)}\alpha|D\ |^{-\frac{1}{2}}\int \exp[-\frac{1}{2\varnothing}\sum_{i=1}^{n}\sum_{j}^{M}\ d_{ij}(\mathrm{y_{ij}};\mathrm{g}(\mu_{ij}^{b}))-\frac{1}{2}b'D^{-1}b]\ db \quad (16)$$

حيث ان :-

$$d_{ij}(\mathrm{y_{ij}};\ \mathrm{g}(\mu_{ij}^{b})=\ -2\frac{y-u}{\mathrm{a_{ij}}v(u)}du$$

ولصعوبة تطبيق الاستدلال الامكان الكامل (الشامل) في التكاملات اللازمة لحساب ql والمشتقات الجزئية الخاصة بها يمكن كتابه المعادلة (16) بالشكل التالي :-

$$\mathrm{c}|D\ |^{-\frac{1}{2}}\int e^{-k(b)}\ d$$

ونلاحظ انه من الصعوبة اجراء الاشتقاق لذلك سنستخدم طريقة لابلاس Laplace' Method لتقريب التكامل، والذي يتطلب مشتقات جزئية مستمرة من الدرجة الثانية نسبه الى b وبعدها نحصل على تقديرات PQL كما يلي :-

ليكن 'k و ''k مشتقات جزئية من المرتبة الاولى والثانية بالأبعاد q و $(\mathrm{q}\times q)$ نسبة الى b، وبتجاهل ثابت المضروب c واخذ اللوغاريتم نحصل على:-

$$ql(\alpha,\theta)\approx-\frac{1}{2}\log|\mathrm{D}|-\frac{1}{2}log\ |\ k''(\hat{b})\ |\ -k'(\hat{b}) \quad (17)$$

إذ أن:

$$\hat{b}=\hat{b}\ (\alpha,\theta)$$

$$k'(b)=-\sum_{i=1,}^{n}\sum_{j=1}^{M}\frac{(\mathrm{y_{ij}}-\mu_{ij}^{b})\mathrm{z_{ij}}}{\varnothing\mathrm{a_{ij}}\mathrm{v}(\mu_{ij}^{b})g'(\mu_{ij}^{b})}+D^{-1}b=0$$

نشتق مرة ثانية بالنسبة ل b:-

$$k''(b)=-\sum_{i=1,}^{n}\sum_{j=1}^{M}\frac{\mathrm{z_{ij}}\mathrm{z_{ij}}^{t}}{\varnothing\mathrm{a_{ij}}\mathrm{v}(\mu_{ij}^{b})\left[g'(\mu_{ij}^{b})\right]^{2}}+D^{-1}+R$$
$$\approx\mathrm{z_{ij}}^{t}wz+D^{-1} \quad (18)$$

و ان:-

D : تمثل مصفوفة التباين والتباين المشترك.

w : مصفوفة قطرية $(\mathrm{n}\times n)$ والتي تعرف بالاوزان التكرارية للانموذج الخطي العام.

$$w_{ij}=\{\varnothing\mathrm{a_{ij}}\mathrm{v}(\mu_{ij}^{b})\left[g'(\mu_{ij}^{b})\right]^{2}\}^{-1}$$

R المدى المتبقي

$$R=-\sum_{i=1}^{n}\sum_{j=1}^{M}\ (\mathrm{y_{ij}}-\mu_{ij}^{b})\ \mathrm{z_{ij}}\frac{\partial}{\partial b}[\frac{1}{\varnothing\mathrm{a_{i}}\mathrm{v}(\mu_{ij}^{b})g'(\mu_{ij}^{b})}]$$

$$E(R)=0$$

وبتجاهل R ودمج المعادلات من (16) الى(18) نحصل على :

$$ql(\alpha,\theta)\approx-\frac{1}{2}\log|\mathrm{I}+Z^{t}WZD\ |\ -\frac{1}{2\varnothing}\sum_{i=1}^{n}\sum_{j=1}^{M}\ (d_{ij}\left(\mathrm{y_{ij}};\hat{\mu}_{ij}^{b}\right)-\frac{1}{2}\hat{b}^{t}D^{-1}\hat{b}\ldots \quad (19)$$

وبافتراض ان الاوزان التكرارية للانموذج الخطي العام قليلة التغيير (او لا تتغير على الاطلاق) كدالة للمتوسط و سوف نتجاهل الحد الاول ونختار α لتعظم الحد الثاني وبالتالي فأن:

$$(\hat{\alpha},\hat{b})=(\hat{\alpha}(\theta),\hat{b}(\theta))$$

حيث ان :-

$$\hat{b}(\theta)=\hat{b}(\hat{\alpha}(\theta))$$

$$\mathrm{Pql}\approx-\frac{1}{2\varnothing}\sum_{i=1}^{n}\sum_{j=1}^{M}\ (d_{i}\left(\mathrm{y_{ij}};\hat{\mu}_{ij}^{b}\right)-\frac{1}{2}\hat{b}^{t}D^{-1}\hat{b}) \quad (20)$$

وبالاشتقاق بالنسبة للـ α و b نحصل على معلمات المتوسط:-

$$\sum_{i=1}^{n}\sum_{j=1}^{M} \frac{\left(y_{ij}-\mu_{ij}^{b}\right)x_{ij}}{\phi a_i v(\mu_i^b)g'(\mu_i^b)} = 0 \qquad (21)$$

$$\sum_{i=1}^{n}\sum_{j=1}^{M} \frac{\left(y_{ij}-\mu_{ij}^{b}\right)z_{ij}}{\phi a_{ij} v(\mu_{ij}^b)g'(\mu_{ij}^b)} = 0 \qquad (22)$$

قام Green في عام 1987 بتطوير خوارزمية فشر الرقمية لحل المعادلتين (21) و(22) على انها مسالة مربعات صغرى موزونة تكرارية (IWLS) تتضمن متغير معتمد ومصفوفة اوزان يتم تحديثها عند كل تكرار وبتعريف المتجه $Y_j = n_{ij}^b + (y_{ij} - \mu_{ij}^b)\, g^t(\mu_{ij}^b)$ فان الحل للمعادلتين (21) و(22) بواسطة خوارزمية فشر يمكن ان يعبر عنه على انه الحل التكراري للنظام كما يلي:

$$\begin{bmatrix} X^tWX & X^tWXD \\ Z^tWX & I+Z^tWZD \end{bmatrix}\begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} X^tWY \\ Z^tWY \end{bmatrix} \qquad (23)$$

وباشتقاق معادلة (23) للحصول على افضل تقدير خطي غير متحيز best linear unbiased estimation (BLUE) للـ α و b في النموذج العادي المرتبط $Y=X\alpha+Zb+\varepsilon$

اذا ان : $\varepsilon\sim N(0, W^{-1})$ و $b\sim N(0, D)$ وان ε و b مستقلان، يكون تقدير التأثيرات الثابته α بالصيغة التالية:-

$$(X^tV^{-1}X)\alpha = X^tV^{-1}Y$$

$$\hat{\alpha} = (X^t V^{-1}X)^{-1}X^tV^{-1}Y \qquad (24)$$

حيث ان:-

$$V=W^{-1} + ZDZ^t$$

وبذلك يكون تقدير b كالاتي:

$$\hat{b} = D\hat{v} = DZ^tV^{-1}(Y - X\hat{\alpha}) \qquad (25)$$

ولتقدير مكونات التباين يتم استخدام معادلة Restricted maximum likelihood (REML) وكما يلي:

$$\frac{1}{2}\left[(Y - X\hat{\alpha})^TV^{-1}\frac{\partial V}{\partial \theta_j}\,(Y - X\alpha) - tr(p\frac{\partial V}{\partial \theta_j})\right] = 0 \qquad (26)$$

اذ ان :-

$$p = V^{-1} - V^{-1}X \times (X^tV^{-1}X)^{-1}X^tV^{-1}$$

تحتوي مصفوفة فشر J على المكونات

$$J_{ij} = \frac{1}{2}\, tr\left(P\frac{\partial V}{\partial \theta_i}p\,\frac{\partial V}{\partial \theta_j}\right)$$

2-4: الامكان شبه الحدي (Marginal Quasi-Likelihood)

تم اقتراح طريقة الامكان الشبه الحدي في عام (1991) من قبل الباحث Goldstein حيث تعتبر هذه طريقة من طرائق التقريبية للاستدلال لتقدير نماذج متعددة المستويات ويتم تنفيذها عن طريق الاستخدام المتكرر للبرنامج القياسي لتحليل مكونات التباين للدوال التي تتبع توزيع طبيعي. لنفرض لدينا مجتمع دراسة معين فعند تقدير التأثيرات العشوائية منفصلة لكل فرد فهذا يعني ان α تمثل تاثيرات متغيره على مستوى الفرد فغالبا من الانسب تحديد النماذج الخطية المعممة من المتوسط الحدي وكما يلي[2]:-

$$E(y_{ij}) = \mu_{ij} = h(x_{ij}^t\,\alpha)$$

ويكتب النموذج بالشكل :-

$$y_{ij} = \mu_{ij}^b + \varepsilon_{ij}$$

$$y_{ij} \approx h(x_{ij}^t\alpha) + h^t(x_{ij}^t\alpha)z_{ij}^t b + \varepsilon_{ij} \qquad (27)$$

وان :-

$$var(\varepsilon_{ij}) = \phi a_{ij} v(\mu_{ij}^b)$$

$$b\sim N(0, D)$$

$$var(y) = V_o + \Delta^{-1}ZDZ^T\Delta^{-1} \qquad (28)$$

وان :-

Δ^{-1} و V_O مصفوفة قطرية عناصرها $\emptyset a_i v(\mu_{ij}^b)$ و $g^t(\mu_{ij})$ على التوالي

نقدر معاملات الانحدار α (التاثيرات الثابتة) باستخدام معادلات شبه الاحتمال كما يلي:-

$$U(\alpha,\theta) = \partial\mu/\partial\alpha^t Var^{-1}(y)(y-\mu) = 0 \qquad (29)$$

حيث ان $\mu = (\mu_1, \dots, \mu_n)^t$ تمثل متجة المتوسط الحدي

$$x^t(\Delta V_O \Delta + ZDZ^t)^{-1}\Delta(y-\mu) = 0 \qquad (30)$$

وباستخدام خوارزمية فيشر الرقمية وكما تم توضيحها في طريقة الامكان الشبه الجزائية PQL التي تؤدي الى انحدار المربعات الصغرى الموزونة التكرارية Iterated weighted least square (IWLS) لمتجة $Y = n + \Delta(y-\mu)$ مع مصفوفة الاوزان:

$$V^{-1} = (\Delta V_o \Delta + ZDZ^t)^{-1} = (W^{-1} + ZDZ^t)^{-1} \qquad (31)$$

حيث ان :-

$n = (n_1, \dots n_n)^t$ تمثل متجة تنبأ الخطي $n_i = x_i^t \alpha$

$w_i = \{\emptyset a_i v(\mu_i)[g^t(\mu_i)]^2\}^{-1}$(عناصر قطرية)

ويتم تقدير التأثيرات العشوائية b وحسب الصيغة التالية:-

$$\hat{b} = DZ^t V^{-1}(Y - X\hat{\alpha}) \qquad (32)$$

حيث ان :-

$\mathrm{b} \sim N(0, D)$

ولتقدير مكونات التباين يتم استخدام معادلة Restricted maximum likelihood (REML) وكما يلي:

$$\frac{1}{2}\left[(Y - X\hat{\alpha})^T V^{-1}\frac{\partial V}{\partial\theta_j}(Y - X\alpha) - tr(p\frac{\partial V}{\partial\theta_j})\right] = 0 \qquad (33)$$

اذ ان :-

$$p = V^{-1} - V^{-1}X \times (X^t V^{-1} X)^{-1} X^t V^{-1}$$

5- دراسة المحاكاة (The simulation study)

1-5: مراحل دراسة المحاكاة (Stages of the simulation study)

لقد تضمنت تجارب المحاكاة استخدام عدد من البرامج وبلغة R في توليد البيانات لغرض المقارنة بين الطرائق باختلاف احجام العينات ولعدة قيم افتراضية للمعلمات .

المرحلة الاولى (The first stage)

في هذه المرحلة يتم اختيار القيم الافتراضية لمعلمات الانموذج والتي تعتمد عليها وبشكل اساسي المراحل اللاحقة وتعد هذه الخطوة من الخطوات المهمة والاساسية والتي تعتمد عليها الخطوات اللاحقة (علما انه تم اختيار قيم المعلمات للبحث من خلال تجربة نفذت على بيانات حقيقية لمرض الجلطة الدماغية والتي تم الحصول عليها من خمسة مستشفيات تابعة لوزارة الصحة العراقية) وكما موضح بالجدول رقم (1) وكما يلي:

جدول رقم (1): القيم الافتراضية لمعلمات الانموذج

Parameters	1	2	3
α_{00}	-2.05537	-1.90000	-2.20000
α_{10}	0.39351	0.50000	0.35000
α_{01}	0.13768	0.16000	0.13200
α_{20}	0.15405	0.18000	0.15000
α_{11}	-0.01736	-0.01400	-0.02100
α_{21}	-0.21238	-0.18000	-0.25000
σ_{00}	0.41327	0.50000	0.37000
σ_{11}	0.86669	0.90000	0.80000
σ_{22}	0.01105	0.01600	0.01000
σ_{01}	.0.23352	-0.20000	-0.25000
σ_{02}	-0.02926	0.03400	0.02600
σ_{12}	-0.03946	-0.03500	-0.04200

المرحلة الثانية (The second stage)

في هذه المرحلة يتم توليد المتغيرات التوضيحية وفق التوزيع الطبيعي Normal distribution وباستعمال الدوال الجاهزة في برنامج (R) حيث تم توليد متغيرين وبالاعتماد على الجانب التطبيقي وكالاتي:-

- $X_1 \sim N(M_1, \sigma_1^2)$
- $X_2 \sim N(M_2, \sigma_2^2)$

يتم توليد التأثيرات العشوائية b_j وفق التوزيع الطبيعي المتعدد Multivariate normal distribution وكما يلي:-

- $b_j \sim MVN(0, D)$

اما بالنسبة الى توليد المتغير المعتمد وفق توزيع برنولي Bernoulli distribution وكما يلي:-

- $y_i \sim ber(m_i)$

المرحلة الثالثة (Third stage)

من اهم العوامل المؤثرة والتي يتم اختيارها هي كالاتي:-
اختيار حجوم عينات مختلفة وهي (n=50,100,150) واختيار تكرار التجربة (1000) مرة.

المرحلة الرابعة (The fourth stage)

في هذه المرحلة يتم تقدير معلمات انموذج الانحدار اللوجستي متعدد المستويات وفق طرائق التقدير والتي تم عرضها بالجانب النظري وهي :-

- Penalized quasi-likelihood
- Marginal quasi-likelihood

المرحلة الخامسة Five stage

في هذه المرحلة تتم مقارنة بين طرائق التقدير المدروسة باستعمال معيار المقارنة متوسط مربعات الخطأ (MSE) للانموذج وكما موضح بالصيغة التالية:-

$$\mathrm{MSE}[\hat{Y}] = \frac{1}{r}\sum_{i=1}^{r}\left\{\frac{1}{n_t}\sum_{j=1}^{n_t}[\hat{Y} - Y]^2\right\}$$

وأن:

r : يمثل مرات تكرار التجربة.

n_t : تمثل حجم العينة لكل تجربة.

2-5: نتائج دراسة المحاكاة (The results of the simulation study)

جدول رقم (2): تقدير معلمات للانموذج الاول وللطريقتين وكافة حجوم العينات

Parameters	N / method	50	100	150
α_{00}	PQL	-2.17540	-2.32387	-1.97795
	MQL	-2.18903	-2.32512	-1.98005
α_{10}	PQL	0.36195	0.43261	0.29954
	MQL	0.38496	0.43499	0.29821
α_{01}	PQL	0.03086	0.20437	0.18444
	MQL	0.03688	0.20597	0.85970
α_{20}	PQL	1.17917	0.27308	-0.23741
	MQL	1.20483	0.27224	-0.23574
α_{11}	PQL	0.07866	-0.02036	0.03958
	MQL	0.08061	-0.02042	0.03935
α_{21}	PQL	-0.64553	-0.27025	-0.10224
	MQL	-0.65307	-0.27141	-0.10348
σ_{00}	PQL	1.61488	0.50635	0.38219
	MQL	1.69834	0.51465	0.38083
σ_{11}	PQL	1.80482	0.87663	0.40816
	MQL	1.93381	0.88312	0.40155
σ_{22}	PQL	1.52048	0.21888	0.12784
	MQL	0.49416	0.21774	0.13206
σ_{01}	PQL	-0.65353	-0.14541	-0.06227
	MQL	0.25242	-0.15246	-0.06551
σ_{02}	PQL	0.94453	0.08433	0.12774
	MQL	-0.09354	0.08790	0.12901
σ_{12}	PQL	-0.55392	-0.02342	0.02073
	MQL	0.23860	-0.02328	0.01782

جدول رقم (3): تقدير معلمات الانموذج الثاني وللطريقتين وكافة حجوم العينات

Parameters	N / method	50	100	150
α_{00}	PQL	-1.90329	-1.82554	-1.84504
	MQL	-1.98223	-1.82622	-1.85189
α_{10}	PQL	0.26077	0.68216	1.00112
	MQL	0.28998	0.68033	1.00577
α_{01}	PQL	-0.00553	0.10852	0.14717
	MQL	0.09528	0.10899	0.14945
α_{20}	PQL	-0.03412	-0.20309	0.09046
	MQL	-0.03410	-0.19735	0.090076
α_{11}	PQL	0.08944	-0.10554	-0.18257
	MQL	0.06177	-0.10536	-0.18535
α_{21}	PQL	-0.18399	-0.05292	-0.17783
	MQL	-0.14107	-0.05368	-0.17873
σ_{00}	PQL	4.95016	0.61814	0.51679
	MQL	2.98470	0.61855	0.51616
σ_{11}	PQL	1.88052	0.75139	0.87017
	MQL	1.59496	0.75229	0.87441
σ_{22}	PQL	1.72089	0.17958	0.13297
	MQL	1.22685	0.18413	0.13483
σ_{01}	PQL	-2.29108	-0.09437	-0.15044
	MQL	-1.48109	-0.09588	-0.14725
σ_{02}	PQL	2.27347	0.02747	0.08219
	MQL	1.32119	0.02284	0.08378
σ_{12}	PQL	-0.94672	-0.01723	-0.00213
	MQL	-0.56314	-0.01467	-0.00018

جدول رقم (4): تقدير معلمات الانموذج الثالث وللطريقتين وكافة حجوم العينات

Parameters	N / method	50	100	150
α_{00}	PQL	-1.97721	-2.52599	-2.59629
	MQL	-2.05518	-2.50049	-2.59741
α_{10}	PQL	-0.80769	-0.21308	0.38540
	MQL	-0.84473	-0.21546	0.39285
α_{01}	PQL	0.00255	0.13325	0.24803
	MQL	0.03526	0.12918	0.24935
α_{20}	PQL	1.42119	0.42827	-0.24794
	MQL	1.45369	0.42719	-0.25974
α_{11}	PQL	0.30504	0.16059	-0.09107
	MQL	0.30373	0.16012	-0.09436
α_{21}	PQL	-0.57783	-0.42301	-0.09512
	MQL	-0.57970	-0.42231	-0.09473
σ_{00}	PQL	1.15757	0.55817	0.31049
	MQL	1.07960	0.55922	0.31181
σ_{11}	PQL	1.56933	0.60007	0.63998
	MQL	1.64740	0.60145	0.64361
σ_{22}	PQL	1.25305	0.27207	0.14821
	MQL	1.30994	0.27044	0.14854
σ_{01}	PQL	0.03310	-0.20902	-0.09457
	MQL	0.03050	-0.21669	-0.09209
σ_{02}	PQL	-0.02233	0.09645	0.00887
	MQL	-0.22401	0.09158	0.00878
σ_{12}	PQL	-0.10710	-0.03527	-0.03461
	MQL	-0.22411	-0.04236	-0.03949

جدول رقم (5):متوسط مربعات الخطا (MSE) للنموذج

Model	N	PQL	MQL	Best
1	50	0.07100393	0.07001598	MQL
	100	0.04343992	0.04343741	MQL
	150	0.03853935	0.03853362	MQL
2	50	0.07094393	0.07072271	MQL
	100	0.05090735	0.05088896	MQL
	150	0.03932679	0.03929627	MQL
3	50	0.06433479	0.06424932	MQL
	100	0.03927800	0.03923398	MQL
	150	0.02952271	0.02950711	MQL

6- الاستنتاجات(Conclusion)

تعد النماذج متعددة المستويات من اهم النماذج الواسعة الاستعمال في تطبيق وتحليل البيانات التي تتصف بكون المشاهدات تتخذ الشكل الهرمي ، وفي بحثنا هذا تم تناول انموذج الانحدار اللوجستي الثنائي متعدد المستويات (انموذج حد ثابت عشوائي وميل عشوائي) , حيث يهدف البحث الى تقدير معلمات هذا الانموذج باستعمال طريقة الامكان الشبة الجزائية (Penalized quasi-likelihood) و طريقة الامكان الشبه الحدي (Marginal quasi-likelihood) ، ومن خلال ما تم طرحه في الجانب التجريبي فقد تم التوصل الى عدد من الاستنتاجات حيث اظهرت نتائج المحاكاة بان قيم متوسط مربعات الخطأ متقاربة جدا للطريقتين ولكافة الحجوم ، وايضا اظهرت نتائج المحكاة بان طريقة MQL هي الافضل لكافة حجوم العينات وذلك لحصولها على اقل قيمة لمعيار متوسط مربعات الخطأ MSE مقارنة مع طريقة PQL ولكافة النماذج , وتم التوصل الى ان الانموذج الثالث والذي استخدمت فيه القيم (-0.2500,-0.370,0.800,0.010,2.250-,0.0260,0.042 -2.200,0.350,0,1320,0.150,0.0210-) هو الافضل وذلك لحصوله على اقل قيمة لمتوسط مربعات الخطأ ولكافة حجوم العينات.

المصادر

1.Al-Zarqani, A. H., (2015), "estimate parameters multiple logistic regression model in the case of a separation problem and multicollinearity with practical application in the medical field", Mcs, university of baghdad.

2.Breslow, N. E. & Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models", Journal of the American Statistical Association, Vol. 88, No. 421, pp. 9-25.1

3.Callens, M. & Croux, C. (2005), "Performance of likelihood-based estimation methods for multilevel binary regression models", Journal of Statistical Computation and Simulation, Vol. 75, No. 12, PP. 1003–1017.2

4.Cook, D., Dixon, P., Duckworth, W. M., Kaiser, M. S., Koehler, K., Meeker, W. Q. & Stephenson, W. R. (2001),"Binary Response and Logistic Regression Analysis", University NSF/ILI project Beyond Traditional Statistical Methods.3

5.Deana, C. B. & Ugarteb, M. D. & Militinob, A. F. (2004), "Penalized quasi-likelihood with spatially correlated data", Computational Statistics & Data Analysis 45, PP. 235–248.4

6.Fagbamigbe, A. F. & Bakre, B. B. (2018), "Evaluating Likelihood Estimation Methods in Multilevel Analysis of Clustered Survey Data", African Journal of Applied Statistics Vol. 5 (1), PP. 351–376.5

7.Hox, j. j., (2010), "Multilevel analysis", second edition ed. utrecht university: the netherland.
8.Handayani1, D., Notodiputro, K. A., Sadik, K. & Kurnia , A. (2017), "A Comparative Study of Approximation Methods for Maximum Likelihood Estimation in Generalized Linear Mixed Models (GLMM)", AIP Conference Proceedings1827,020033 , https://doi.org/10.1063/1.4979449, PP.1-10.6
9.Khan, M. d. & Shaw, J. H. (2011), "Multilevel Logistic Regression Analysis Applied to Binary Contraceptive Prevalence Data", Journal of Data Science 9, pp. 93-110.7
10. Moineddin, R., Matheson, F.I. & Richard, R. H. (2007), "A simulation study of sample size for multilevel logistic regression models", BMC Medical Researc Methodology,vol7, no.34, http://www.biomedcentral.com/1471-2288/7/34, PP.1-10.8
11. Shahmandi, M., Farmanesh, F., Gharahbeigi, M. M. & Shahmandi, L. (2013), "Data Analyzing by Attention to Weighted Multicollinearity in Logistic Regression Applicable in Industrial Data", British Journal of Applied Science & Technology 3(4), pp. 748-763.9
12. Sommet, N & Morselli, D. (2017), "Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS", International Review of Social Psychology, 30(1), 203–218, DOI: https://doi.org/10.5334/irsp.90.10
13. Moerbeek, M., Van Breukelen, G. J., & Berger, M. P. (2003), "A comparison of estimation methods for multilevel logistic models", ***Computational Statistics*****, 18(1), 19-37.11**

Comparison between method penalized quasi- likelihood and Marginal quasi-likelihood in estimating parameters of the multilevel binary model

Zainab Nihad Mohammed ALrawi
University of Baghdad / Computer Center
zainab.alrawi@yahoo.com

Asst.prof.Iman Hasan Alani
/ University of Baghdad/College of Administration & Economics /statictics Department
eman.alani1962@gmail.com

Received:5/5/2020 Accepted :21/6/2020 Published :October / 2020

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

Abstract:

Multilevel models are among the most important models widely used in the application and analysis of data that are characterized by the fact that observations take a hierarchical form, In our research we examined the multilevel logistic regression model (intercept random and slope random model) , here the importance of the research highlights that the usual regression models calculate the total variance of the model and its inability to read variance and variations between levels ,however in the case of multi-level regression models, the calculation of the total variance is inaccurate and therefore these models calculate the variations for each level of the model, Where the research aims to estimate the parameters of this model using approximation methods (penalized quasi- likelihood and Marginal quasi- likelihood), A simulation method was used to compare the estimation methods for different sample sizes, through Mean squared error (MSE) to get the best method to estimate the parameters, the result obtained using the simulation method showed that the estimation methods gave close result, but the method (MQL) is the best in all sizes .

Keywords: Logistic Regression, Laplace Method, Semi - Integrated Likelihood Function, Digital Fisher Algorithm, Iterated Weighted Least Square.